LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-469012

Viral shift and drift, report part 1

M. Borucki, J. Allen, T. Slezak

February 4, 2011

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-469012

Viral shift and drift, report part 1

Investigators

Jonathan Allen, (925) 422-0662, allen99@llnl.gov

Monica Borucki, (925) 424-4251, borucki2@llnl.gov

Thomas Slezak, (925) 422-5746, slezak1@llnl.gov

Lawrence Livermore National Laboratory

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Viral shift and drift, report part 1

Jonathan E. Allen, Monica K. Borucki, Tom R. Slezak Lawrence Livermore National Laboratory

The ability for RNA viruses to evolve at a high rate has the potential to confound existing medical technology by making it more difficult to accurately diagnosis and treat RNA based infectious disease. The rapid mutation rates of RNA viruses give a virus the ability to adapt to novel selective pressures, which can potentially support zoonosis, sustained outbreaks in a novel host, and resistance to anti-retroviral treatments. A key contributor to an RNA virus' mutation rate is the lack of a proof reading enzyme to correct for mis-incorporated bases when transcribing a daughter RNA particle. Thus, there are potentially many opportunities for new mutations to be randomly introduced and subsequently fixed in a key subset of the viral population. This is a feature, which we refer to in this report as genetic drift. One of the hurdles to understanding the role of genetic drift in the spread of RNA virus outbreaks is the need to analyze a densely sampled outbreak using whole genome sequencing in order to measure the amount of mutation occurring on a small time scale. The recent advances in sequencing technology have now lowered the data collection cost barriers to allow the measure of evolution not just among the consensus virus sequences on a shorter time scale, but potentially capture all the distinct genetic variants that are circulating within a single infected host. Tracking the distinct genetic variants within an infected individual provides additional information on genetic drift within the host, and can help determine whether mutations that are important for emerging disease are able to persist at the sub consensus level.

We thus are undertaking an effort to identify the benefits and challenges to measuring genetic drift and its role in viral outbreaks through the examination of a case study, a rabies virus outbreak in Northern California. In collaboration with the California Department of Public Health, we obtained access to 50 samples of rabies infected animals (primarily foxes, and skunks). In order to gain an understanding of how the sequencing technology can be used to measure small numbers of genetic mutants in a population and determine the degree of sensitivity to detect variation needed, we sequenced three samples with "ultra-high" sequencing coverage using an Illumina sequencer. The results of this experimental work are the focus of this report, which generates the data that is used to devise a sequencing strategy for the remaining samples. Here is a brief listing of the project results detailed in this report, each of which will be explained in the following sections. Results include:

- A rare variant SNP detection pipeline for Illumina data.
- An estimate for the total number of mutations within a single fox infected host.
- An estimate for the relative frequency of mutations within a single fox infected host.
- Comparison of mutants in the two infected fox samples.
- A practical strategy for sequencing a large number of samples from a viral outbreak.

In the initial experiment, our goal was to evaluate the potential for the latest high throughput sequencing tools to detect genetic variants within a single host including rare variants and determine how much sequencing coverage is needed and how best to overcome the inherent noised introduced from DNA amplification and sequencing errors. Three samples were sequenced each in separate lanes of a single flow cell of an Illumina IIx sequencer using paired-end reads on short genomic fragment inserts using read lengths of approximately 112 bases. The three samples consisted of a plasmid control containing a 1 kb insert for the rabies virus polymerase gene. Two additional samples were taken from brain tissue of rabies-infected foxes from the Northern California outbreak. All samples were amplified using PCR prior to sequencing. To capture the whole genome of the rabies overlapping PCR amplicons were generated that span the length of genome. Since the PCR primers could potentially introduce false mutations into the amplicon pool due to none specific binding, all of the primer regions were masked out for the downstream analysis. Table 1 summarizes the output generated in the initial runs.

Table 1 Data generated for initial Illumina sequencing run.

Sample	Sequence output (gigabases)	Number of Reads
Plasmid control	6.3	2 x 28,325,049
Rabies 1	6.06	2 x 27,063,566
Rabies 2	6.06	2 x 27,051,934

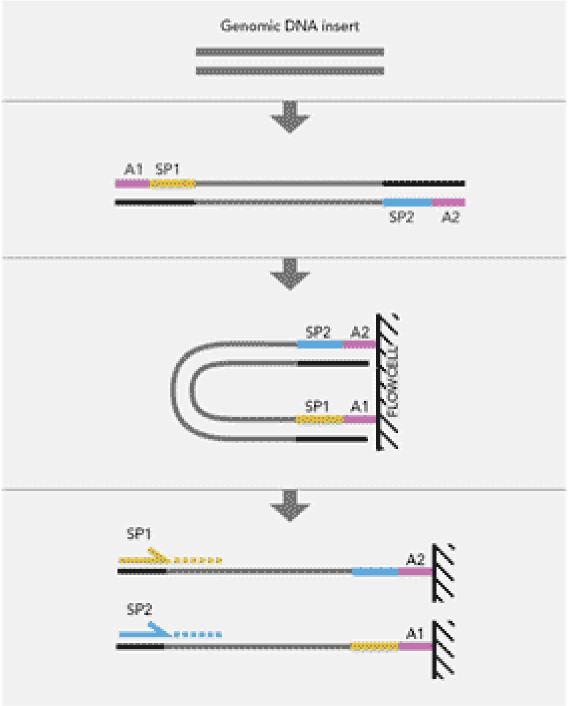


Figure 1. Process for sequencing paired-end reads. (Image taken from Illumina.com)

Figure 1 shows Illumina's procedure for generating a single "paired-read", where the genomic fragment is placed on the sequencer's flowcell prior to clonal amplification. Effective genome coverage is reported later in the report as the average number of distinct sequenced DNA inserts covering any given position in the genome. The raw coverage value measuring all mapped reads was approximately 450,000x coverage. It must be noted that calculating effective coverage (later in the report) on a per sequenced insert basis rather than on an individual per read basis, avoids "double counting" of overlapping reads taken from the same insert. Figure 2 shows a schematic of how the two paired reads overlap given the observed average insert length of 142, with a range of from 95 to 188 normally distributed (personal communication Eureka genomics).

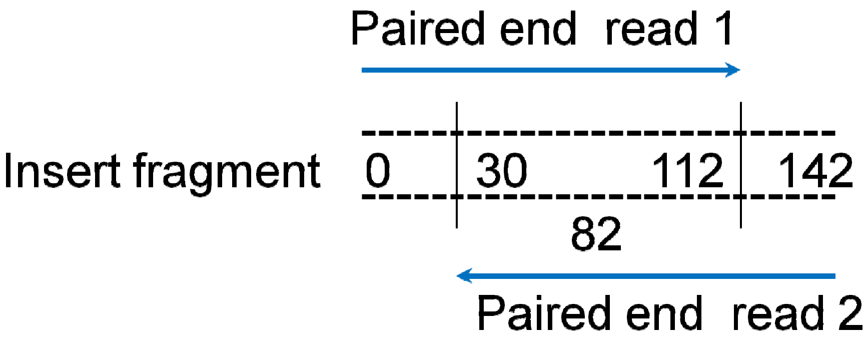


Figure 2. Example of paired-end read coverage of a sequenced insert, where the read lengths are 112 and the observed average fragment length of 142. Dotted lines denoted the double stranded DNA insert fragment, numbers between the dotted lines denote positions in the sequence (from 0 to 142). The diagram shows an 82 base overlap between the two reads.

A rare variant SNP detection pipeline

The plasmid control sample was used to empirically model potential false SNP calls introduced through errors generated from the PCR amplification of the sample and errors introduced by the sequencing by synthesis reactions carried out by the Illumina sequencer (Fuller et al., 2009). Thus, the initial single clone control sample was amplified using the same PCR amplification protocol and sequenced. The control reference sequence generated with a separate Sanger sequencing run and the sequenced reads were mapped using our standard read mapping protocol, which is applied identically for all sequenced samples. Any polymorphisms that deviate from the consensus sequence are taken to be examples of error introduced at either the PCR amplification step or the sequencing step. Figure 3 shows the procedure for computing an observed error rate, by calculating the percentage of non-consensus base calls at each position in the reference genome from genome position 0 to N to produce a distribution of observed error values that effectively measure the expected number of miscalls at any given position. Note that since the samples of interest are RNA viruses, there is an initial reverse transcription step, where errors can occur but cannot be modeled, thus the error rate of the reverse transcription step presents a fundamental limit on the ability to distinguish experimental errors from real mutations within the sample's viral population. Given the relatively low error rates of the reverse transcriptase and the lack of clonal amplification in this implies, that these types of error will likely make up a very small percentage of the overall error.

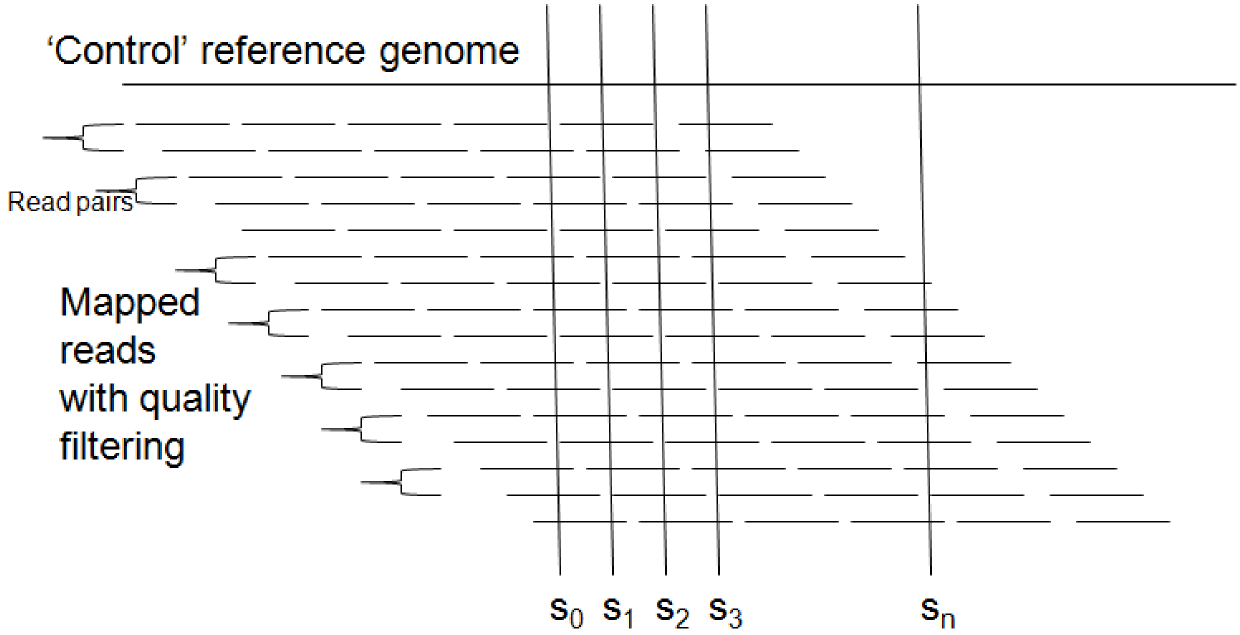

'Control' reference genome
Read pairs
Mapped reads with quality filtering
s_0
s_1
s_2
s_3
s_n

Figure 3. Procedure for calculating observed error rates. Each position in the genome, S_0 to S_N is treated as an independant observation, where the percentage of miscalls are tallied to give N observed error rates.

The read mapping pipeline applies an open source read mapping software tool, SHRiMP (Rumble et al. 2009), which was chosen for the tool's ability to conduct sensitive mapping (using an optimal Smith-Waterman alignment) to map as many reads as possible in the face of individual errors within each read. The goal is to map as many reads as possible, and then carefully evaluate each read's potential to reliably contribute evidence for a variant in the population, depending on the specific variant in question. Thus, a key requirement is that the reads be mapped to a reference sequence, and the SNP calling procedure involves enumerating over every position in the reference genome to consider the possibility of a nucleotide variation occurring at that position. Hence, novel base insertions are not considered. Note that although a reference sequence is used, this is not expected to be a limitation in cases where an existing reference sequence does not exist. Given the relatively small genome size of the RNA viruses and lack of large scale genome rearrangements a consensus reference sequence should be recoverable from the raw reads. Moreover, since the protocol uses a PCR amplification step it is likely whatever original reference sequences were used to design the amplification primers, likely will not be too divergent in practice. When evaluating each position in the reference genome, a simple rule set is applied to decide when a read should contribute to the presence or absence of a variant: the read base call quality score must be 35 or higher (an observed above average quality score) over an 11 nucleotide window (+-5 bases around the query nucleotide). Any predicted indel is excluded from consideration. A second additional feature was evaluated, which was to require that read pairs overlap the query region and agree on the same base call.

Using paired-end reads improves Illumina ultra-rare variant detection

Figure 4 shows the analysis of the control sequence based error rates comparing the use of each read independently (blue line) versus using only the overlapping paired reads with agreeing base calls (green line). The x-axis gives the relative frequency of observed error rate across the control sequence and shows that two distinct error models emerge. On average, the error rate for the paired read approach is about half the rate of using reads independently (0.00025 versus 0.0005) but equally importantly, the range of error rates is much smaller for the paired end data, with a maximal observed error rate of 0.00058. By contrast relying on single read derived base calls introduces higher variance. Thus, to make high confidence SNP calls using single read derived calls must in practice assume a much higher error rate to preclude the possibility of errors that are less frequently introduced but still substantial in number.

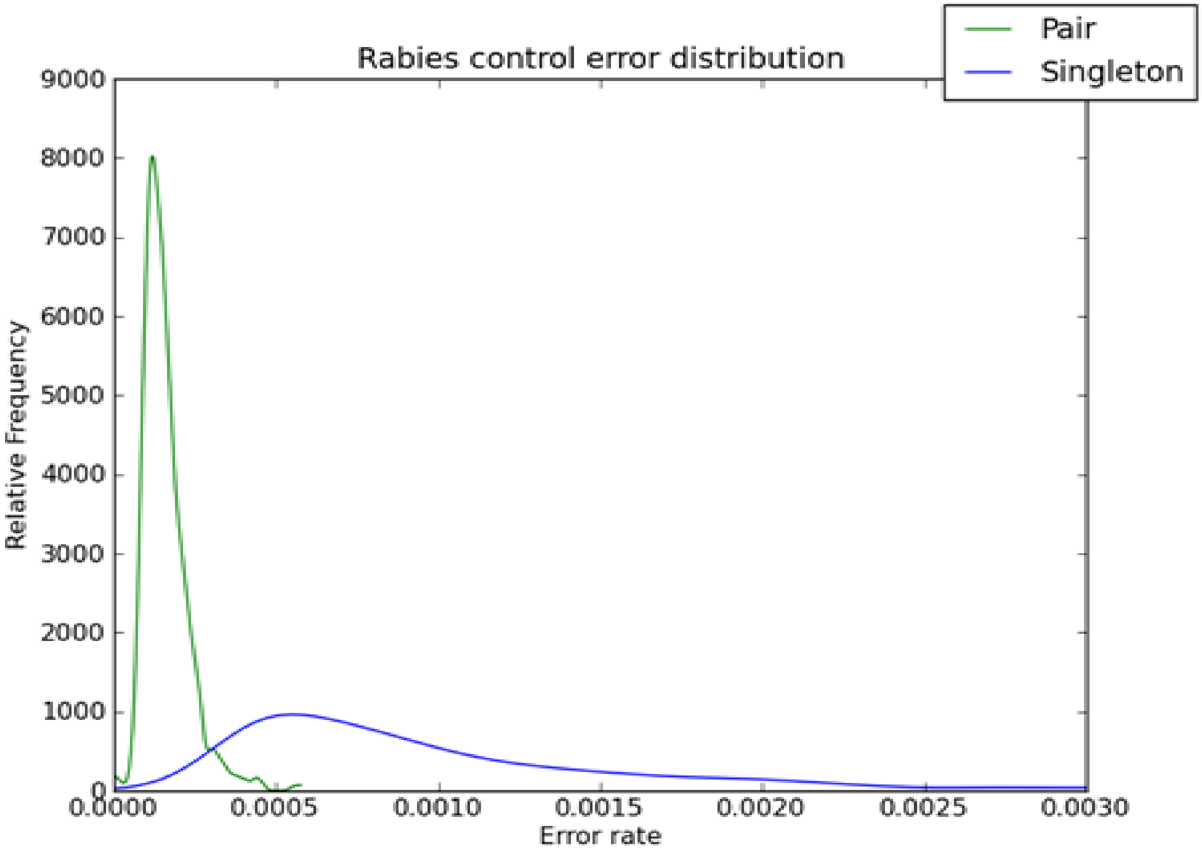
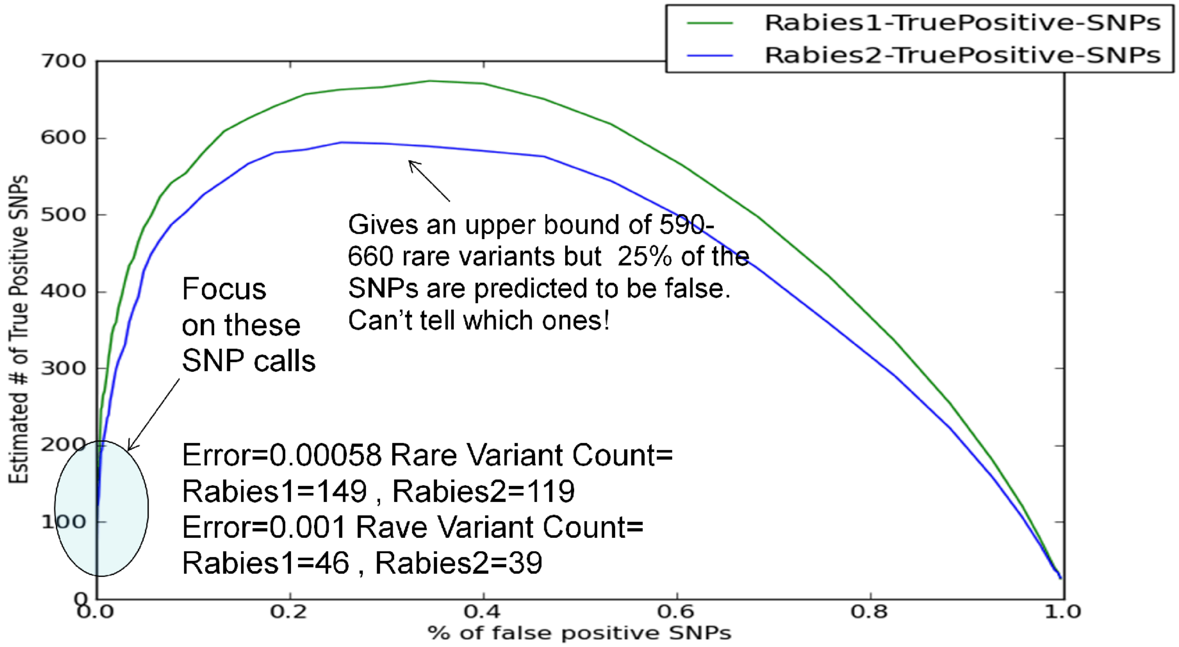


Figure 4. Shows a probability density plot for the two types of error models, using each read independently (blue) or using paired reads for error validation (green).

Currently the error correction approach described by Eriksson et al. is used, which defines a Binomial error model in terms of the observed error rate and the total number of overlapping reads. The model defines the expected number of non consensus bases to occur due to random error given the error rate for a given number of observed reads, using a preset P-value. In order for a non-consensus base call to be made, a minimum number of reads must occur that exceed the amount that is expected by random chance given the error rate and sequencing coverage. The P-value was set to 0.01, since the experiments will look typically at 10,000 positions in the genome, this is the value used to correct for multiple hypothesis testing.

An estimate for the total number of mutations within a single host



Gives an upper bound of 590-660 rare variants but 25% of the SNPs are predicted to be false. Can't tell which ones!

Focus on these SNP calls

Error=0.00058 Rare Variant Count=
Rabies1=149 , Rabies2=119
Error=0.001 Rave Variant Count=
Rabies1=46 , Rabies2=39

Binomial test conditioned on different error rates

Figure 5. Estimate for the number of mutants present in the population, conditioned on different error rates. Y-axis shows different estimates of total SNP counts, versus the percentage of predicted SNPs from the total pool that are expected to be false positives. The number of high confidence SNPs are highlighted where the x-axis is at 0.

Figure 5 shows an estimate for the expected number of mutations using different confidence thresholds on the error rates using the paired end read error model shown in Figure 4. The idea is that the true probability of an error at any given position in the genome is governed by the distribution in Figure 4, thus the most conservative estimate uses the highest observed error rate (0.00058), but the distribution suggests that this value may be too strict and lead to false negatives. To provide additional estimates of the total true SNP count, the error rate is iteratively dropped, but as the error rate goes down the probability that the true error rate is in fact higher (leading to false positives) goes up and must be taken into account. For example, if the presumed error rate is 0.00029, according to the distribution, there is a 5% chance that the true error rate is in fact higher. Therefore, when using 0.00029 as the error rate in the Binomial test, in the example of Rabies 1 sample would predict 508 SNP calls, but we expect 25 (5%) of these to be false positives, to account for when the true error rate is higher. Using this approach suggests that the maximum number of SNPs for the rabies 1 and rabies 2 samples are 673 and 593 respectively. This raw count is drawn from a much larger pool of candidate SNPs, for which we know many must be false positives, and the difficulty lies in the fact there is no empirical way to distinguish the pool of real variants from errors. Instead, we focus on the "high confidence" SNPs, which assumes a "worst case" higher error rate, and thus require that more reads supporting the mutant must be present before reporting the mutation is real.

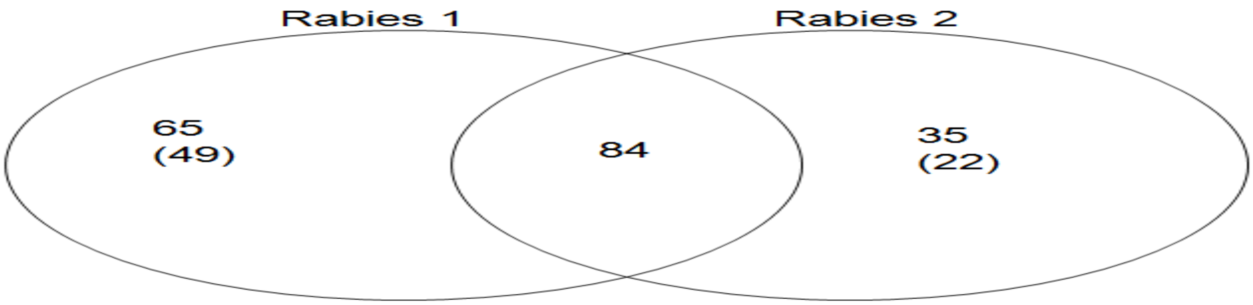


Figure 6. Shared and unique high confidence mutants in both rabies populations. 84 shared SNPs. Number in parentheses show number of SNPs where the regional coverage is higher in the sample containing the SNP call.

Quality control is critical to prevent "over-interpreting" biological results

Figure 6, shows the breakdown of high confidence mutations between the two rabies samples, with 84 shared between the two samples, and Rabies 1 and Rabies 2 having 65 and 35 of their own respectively unique mutant set. The fact that the rabies 1 sample shows more mutants than the rabies 2 sample highlights an important factor of the sequencing process, which is the potential lack of uniformity of coverage across multiple samples. With higher coverage comes the potential to report more rare variants, thus it is important when comparing multiple samples to determine whether the reported presence or absence of a mutant relative to comparable samples is not due to higher or lower coverage. In the majority of cases 49 of 65 for rabies 1 and 22 of 35 for rabies 2, cases where a sample specific mutant is reported, correspond to rare variants, where the coverage is also higher in that sample. Moreover, interestingly, there was noticeable difference in the quality of the sequencing output generated for the two samples, which is highlighted in Figure 7.

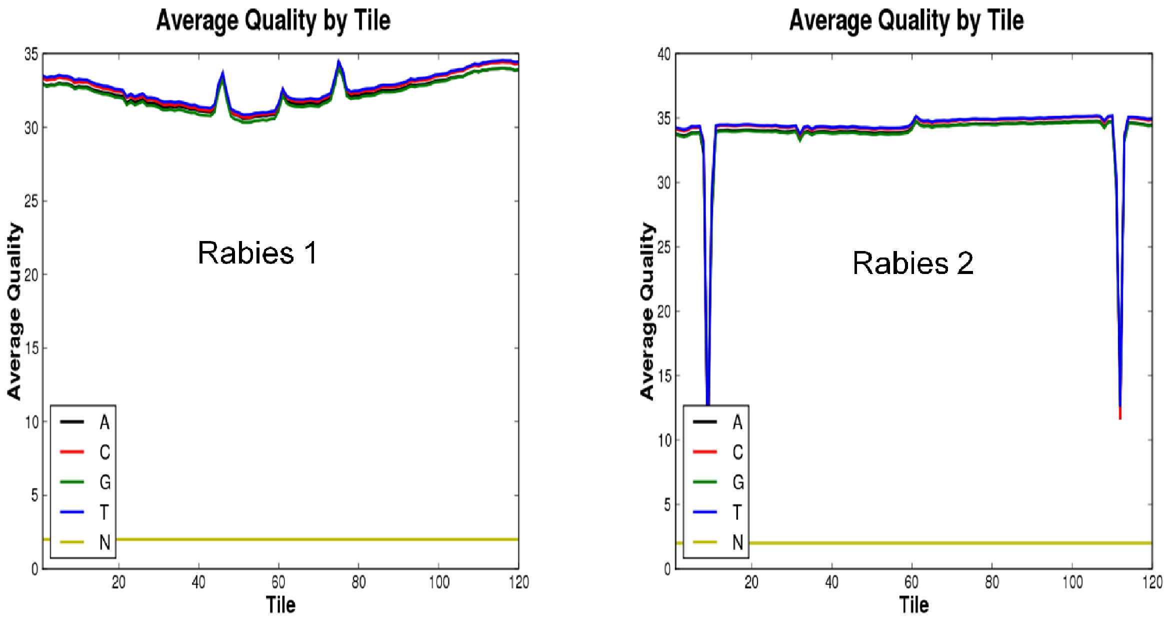


Figure 7. Average quality scores across the physical layout of the flow cell. (Figure taken from Eureka genomics quality control report.)

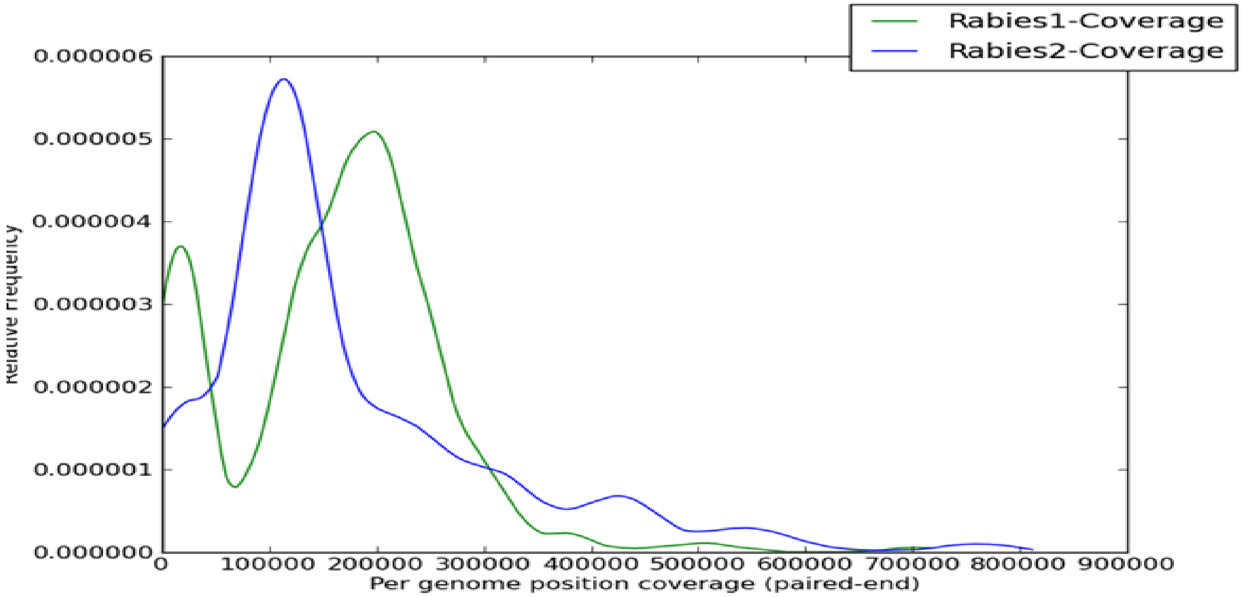


Figure 8. Histogram showing the distribution of effective coverage in each rabies sample.

The figure compares the average quality scores for sequenced reads and highlights the fact that there are two regions in the flow cell, where the quality scores of the reads of the rabies 2 sample dip far below the average. The pipeline's quality filtering will thus ignore many reads with the lower quality scores leading to a lower effective coverage for the rabies 2 sample. This observation is confirmed by Figure 8, which shows that the most common effective coverage (after quality filtering) exceeds 200,000x coverage for the rabies 1 sample, but is closer to

100,000x coverage for the rabies 2 sample. Thus, the difference in mutant count between the rabies 1 and rabies 2 is explained not by a biological difference, but through inherent differences in the sequencing process.

An estimate for the relative frequency of mutants within a single fox infected host.

The majority of the identified mutants are exceedingly rare in the population. Even when applying the more conservative error rate, the high 100,000x+ coverage allows for exceptionally sensitive detection of rare variants down to 0.08% level. For the purpose of this work, sensitivity is defined in terms of the number of reads showing the variant divided by the total number of reads overlapping the genome position. Thus, a variant that occurs with 0.08% frequency at a region of 100,000x coverage would imply 80 reads contain the variant. Figure 9 shows the cumulative distribution for the mutant frequency in the population for the two samples. The figure indicates that 90% (Rabies 1) and 92% (Rabies 2) of the mutations occur with less than 1% frequency and roughly 80% of mutants occur at a frequency of less than 0.2% frequency.

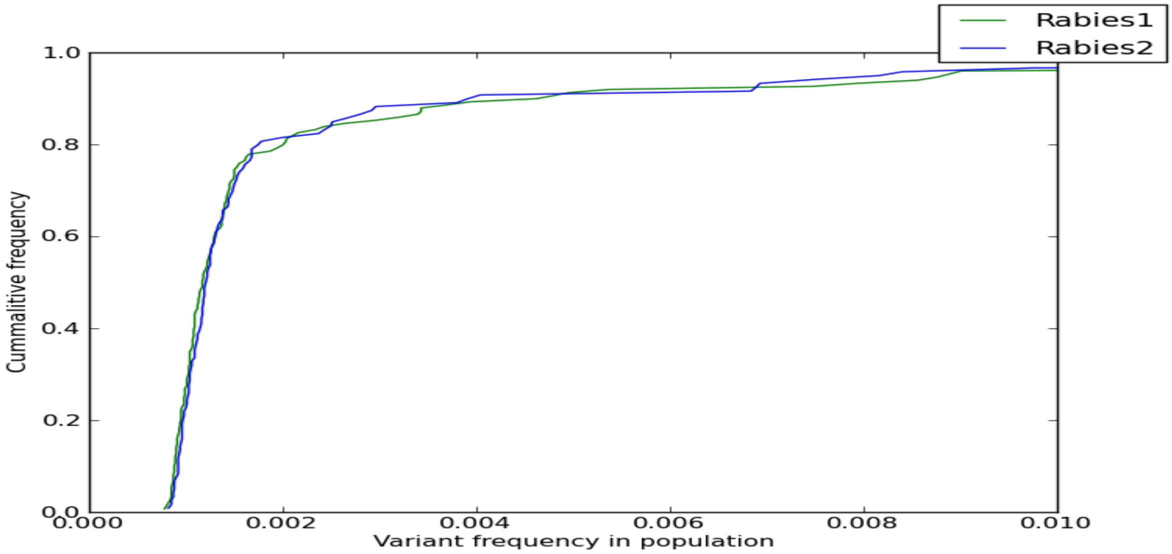


Figure 9. Within host population diversity. X-axis shows relative frequency within the population and y-axis shows the percentage of mutants in the population that occur at this frequency.

Hypothesis proposal: ultra-rare variants have limited functional significance

With such a large percentage of the rare variants occurring at the ultra rare (sub 0.2%) level, the biological significance of these observations remain in question. To get a very rough estimate for the potential significance of the rare variants, the two rabies samples were compared and shared mutants between the two samples were plotted according to their average relative frequency in their respective populations. These values were compared with the relative frequency of mutants that show up in just one sample but not the other. The results are shown in Figure 10. The mutations shared by samples are hypothesized to variants that may be persistent within the Northern California outbreak, while sample specific mutants are hypothesized to be "transient", that is, random population snapshots that include viral replicons with either limited or no fitness and are detected as a consequence of the ultra sensitive sequencing method.

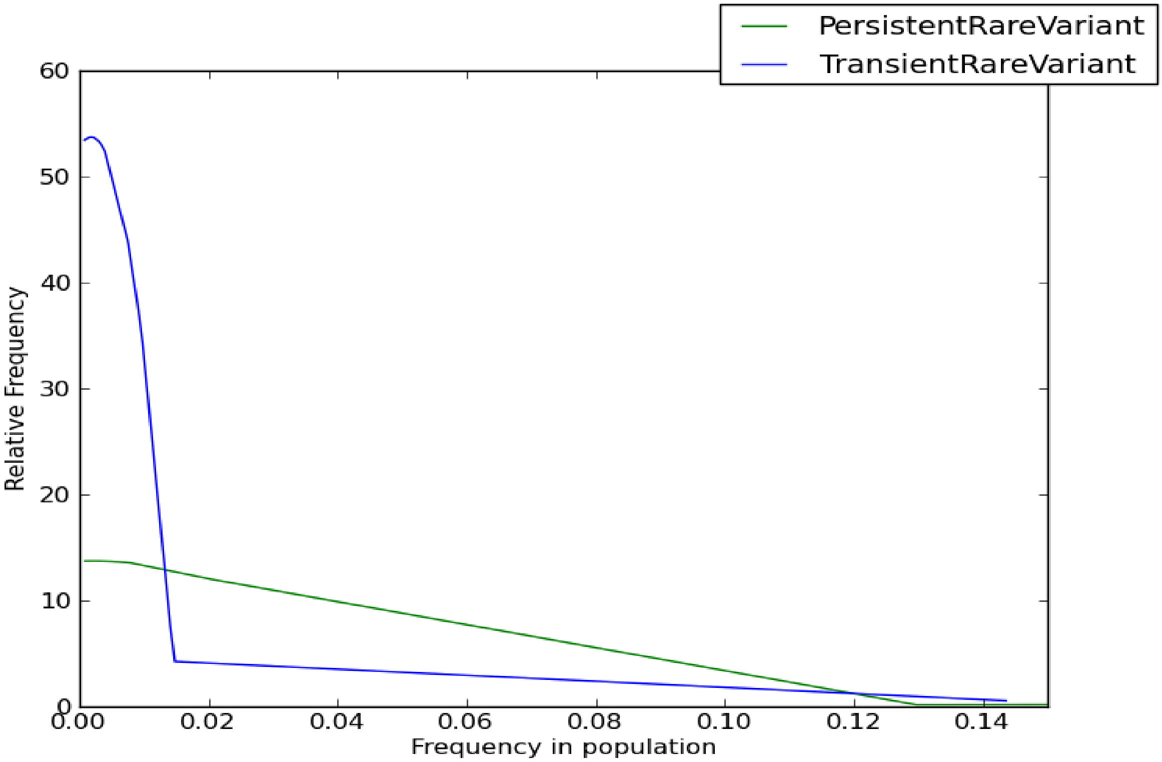


Figure 10. Relative frequency within the population of mutants shared by both samples (PersistentareVariant) versus mutants ahat occur in just one of thetwo samples (TransientRareVariant).

Figure 10 shows nearly all of the so-called transient mutants are concentrated at the ultra-rare level, whereas there is a wider distribution of frequency values for the persistent mutants, that includes both ultra-rare and more common variants. While this is just a limited sample size of two it is tempting to speculate that this provides support to the hypothesis that there is an enrichment for non-functional random variation at the ultra-rare variant level. This data may ultimately provide additional information for better understanding the raw mutation rates and shed light on the potential for random mutations to rise to prominence due to random drift.

A practical strategy for sequencing a large number of samples in a viral outbreak

In the next steps of the project, the sequence data will be expanded from the current sample count of 2 to 41. It will not be practical with respect to managing financial resources to dedicate ultra high levels of sequencing coverage to every viral sample. Moreover, our results show that it may not be particularly useful to pick up every random variant. The ultra high coverage experiments can be used to provide a basis for determining the level of variation that may be present, and what can be detected, and thus give a principled basis for devising a "scaled back" strategy designed to sequence larger numbers of viral samples.

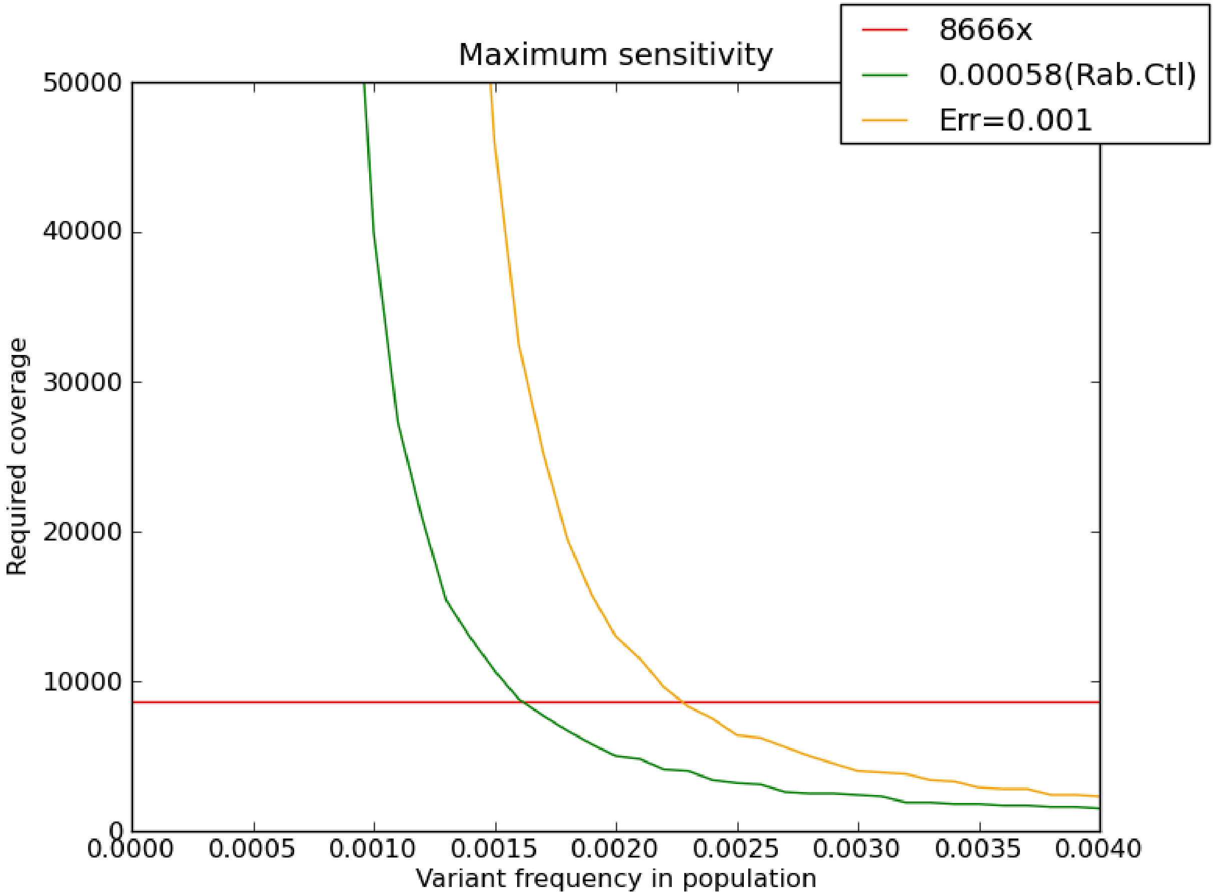


Figure 11. Sequencing coverage requirements for characterizing mutant spectra in a viral population.

Our proposed approach is to use the maximal number of samples (12) per sequencing lane currently supported by Illumina such that each sample can be tagged with a barcode, for unambiguous separation of samples within a single sequencing run. The process effectively reduces both the cost by an order of 1/12 (additional per sample library costs arise) as well as the coverage. Using the observed error rates from our initial sequencing run provides the basis for anticipating the impact of the reduced coverage on the ability to fully characterize the mutant spectra within the sampled population. Figure 11 shows the anticipated minimum frequency at which a mutant can occur within the sample of reads (x-axis) as a function of total coverage (y-axis). Two anticipated error rates are shown, the observed error rate from the sequencing control and a higher more conservative error estimate of 0.001. One important observation to note is applying the rigorous statistical thresholds leads to exponential increases in sequence coverage required for small increases in sensitivity. For example, with an error rate of 0.00058, the amount of coverage required to increase sensitivity from 0.0015 to 0.001 requires an increase from roughly 10,000x coverage to 50,000x coverage. It is important to note that for practical purposes, the limit of sensitivity is determined by the error rate rather than sequence coverage. This further underscores the importance of using the paired end sequencing reads to lower error rate, even at the expense of potentially reduced coverage.

Assuming that each Illumina lane will replicate current preserved performance, 26 million paired end reads can be generated (per lane). Assuming a conservative 75 bases of usable read length and coverage of the 12 kb genome yields a raw coverage level of 13,541x when dividing each lane into 12 separate barcoded samples. Moreover, based on observation of assuming roughly 80% of the reads being usable for base calling and accounting for a 20% variation in coverage due to the fact that each barcoded sample is not sequenced at perfectly equal amounts yields a lower bound coverage estimate per sample of 8,666x coverage. The red line in Figure 11, highlights the anticipated minimum sensitivity for detecting rare mutants in the population for the two distinct error rates suggesting that sensitivity will be reduced by half from 0.08% (using the ultra high 100,000x+ coverage) to 0.16%. The current estimated total per cost sample for this process is approximately $1,200 and we are proceeding with this approach for sequencing the remaining 39 rabies samples.

References

Eriksson N, Pachter L, Mitsuya Y, Rhee S-Y, Wang C, et al. (2008) Viral Population Estimation Using Pyrosequencing. PLoS Comput Biol 4(5): e1000074.

Fuller, C. W.; Middendorf, L. R.; Benner, S. A.; Church, G. M.; Harris, T.; Huang, X.; Jovanovich, S. B.; Nelson, J. R.; Schloss, J. A.; Schwartz, D. C. & Vezenov, D. V. The challenges of sequencing by synthesis Nature Biotechnology, 2009, 27, 1013-1023

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, et al. (2009) SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Comput Biol 5(5): e1000386. doi:10.1371/journal.pcbi.1000386